DOCUMENT RESUME

ED 211 594 TM 820 029

AUTHOR Gustafsson, Jan-Eric
TITLE An Introduction to Rasch's Measurement Model.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.; National Swedish Board of Education, Stockholm.; Swedish Council for Research in the Humanities and Social Sciences, Stockholm.
REPORT NO ERIC-TM-79
PUB DATE 80
NOTE 47p.; Paper presented at the Nordic Researchers' Course "Rasch models in the social and behavioral sciences" (September 29-October 6, 1980).
AVAILABLE FROM ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, NJ 08541 ($5.50).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Ability; Academic Achievement; *Difficulty Level; *Latent Trait Theory; Mathematical Models; *Measurement Techniques; *Test Items
IDENTIFIERS *Rasch Model

ABSTRACT

Some basic concepts of the one-parameter logistic latent-trait model, or the Rasch model, are presented. This model assumes that the probability of a correct answer to an item is a function of two parameters, one representing the difficulty of the item and one representing the ability of the subject. The purpose of this paper is to explain a mathematical-statistical solution to the problem of separating the factor of item difficulty from person ability. Rasch's "theory of specific objectivity" is basically a theory of unidimensionality. It states that only one dimension should be measured at a time and all items should be homogeneous, or measure the same ability. Local stochastic independence is assumed. The Rasch model is compared to other test-theoretic models including classical test-theory and other item-response models. In relation to a concrete example, it is demonstrated how the parameters in the model can be estimated and how the assumptions of the model can be tested. Possible areas of application of the Rasch model are discussed. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED211594

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

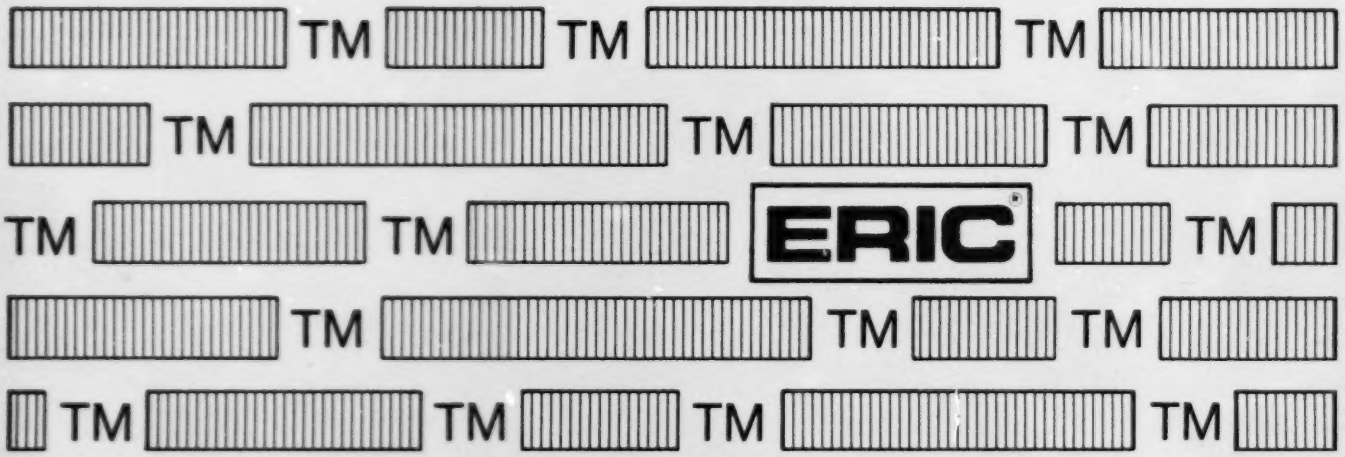
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AN INTRODUCTION TO RASCH'S MEASUREMENT MODEL

ERIC/TM REPORT 79

by
Jan-Eric Gustafsson



TM 820 029

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08541

ERIC/TM Report 79

An Introduction to Rasch's Measurement Model

by

Jan-Eric Gustafsson

University of Gothenburg

Paper presented at the Nordic Researchers' Course
"Rasch models in the social and behavioral sciences,"
September 29 - October 6, 1980.

(The research presented in this paper has been supported financially by the Swedish Council for Research in the Humanities and Social Sciences and by the National Board of Education.)

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service, Princeton, New Jersey 08541

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to qualified professionals for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view of either these reviewers or the National Institute of Education.

ERIC Clearinghouse on Tests, Measurement,
and Evaluation
Educational Testing Service
Princeton, NJ 08541

September 1981

ABSTRACT

The paper presents some basic concepts of the one-parameter logistic latent-trait model, or the Rasch model. This model assumes that the probability of a correct answer to an item is a function of two parameters, one representing the difficulty of the item and one representing the ability of the person. In relation to a concrete example, it is demonstrated how the parameters in the model can be estimated and how the assumptions of the model can be tested. Some possible areas of application of the Rasch model are also discussed.

INTRODUCTION

It is often the case in educational and psychological measurement that we want to assess the same dimension by using different items for different groups of persons. For example, we may want to measure educational achievement several consecutive years without risking that the test items become known; or the level of performance may be much lower in one group of persons than in another group of persons, so that the same items are not suitable.

A concrete example is presented below. Two three-item tests have been administered to two samples of persons (Sample 1 and Sample 2), each consisting of 10,000 persons. The tests were constructed in such a way that one of the items (item 3) is common to the tests, while all other items are different. The following descriptive statistics were computed for Sample 1:

Item	Proportion correct	
1	0.85	
2	0.50	Mean 1.51
3	0.16	SD 0.79

For Sample 2 the following descriptive statistics were computed:

Item	Proportion correct	
1	0.82	
2	0.56	Mean 1.65
3	0.27	SD 0.86

For item 3 we can, of course, compare the samples directly, and we find that Sample 2 displays a somewhat higher level of performance than Sample 1 (0.27 vs. 0.16). But about the rest of the tests not much can be said. It can be observed, however, that the small mean difference in favor of Sample 2 on the total test is almost completely accounted for by item 3. This does seem to indicate that the test given to Sample 2 was more difficult than the test given to Sample 1, since we otherwise would expect about the same difference in level of performance on the noncommon items as on the common item.

Except for these general statements, it does seem quite impossible, however, to use the descriptive statistics for a fuller comparison of the levels of performance in the two samples. Only if we were able to determine how much more difficult the items given to Sample 2 are, might it be possible to "translate" the scores on this test in such a way that they are comparable to scores on the test given to Sample 1.

It would, at first glance, appear to be quite a difficult task to determine the level of difficulty of a set of items. However, most of the remainder of this paper will be devoted to a demonstration that this is in fact possible, and we will show how our problem of comparing the levels of performance of Samples 1 and 2 can be solved.

The solution is based on the quite simple notion that level of performance on an item is governed by two factors: by the ability of the person taking the item and by the difficulty of the item. While the terms ability and difficulty are quite firmly founded in our common sense, we can achieve a slightly higher precision in these concepts through the following statements:

1. For any given item, a person with a higher ability should have a higher chance of passing the item than a person with a lower ability.
2. For persons with the same level of ability, the chance of passing an easier item should be higher than the chance of passing a more difficult item.

Common sense could easily accept the view that for any given person the probability of answering a certain item correctly should be a function of the person's ability and the item's difficulty, even though other factors may also influence the result. But when we have just observed a certain outcome -- a correctly solved item, say -- problems occur if we want to attribute that result to either of these two factors: The person may have solved the item correctly because of a high level of ability or because the item was easy. Similarly, if we observe that a person fails an item, this may be due to the fact that the item was quite difficult, or it may be that the person has a low ability to solve this kind of item.

Thus, while common sense might accept the concepts "level of difficulty" and "level of ability," it is somewhat mind-staggering to consider the possibility of separating these factors in empirical data. There is, however, a mathematical-statistical solution to this problem, which was originally contributed by Rasch (1960). The purpose of this paper is to present this solution and the accompanying model (the Rasch model) in a simple manner; and, even though a series of algebraic expressions will be presented, it is hoped that the mathematically untrained reader can get at least an intuitive understanding of the basic concepts of this test-theoretic model.

THE RASCH MODEL

In developing our case for how to separate person ability and item difficulty, we will start with an assumption about how these factors relate to the probability of answering an item correctly. We first must introduce some notation, however. In all, there are n persons who have answered k items; we will refer to any specific item as item i and to any specific person as person v. The outcome of the encounter between a person and an item will be referred to as A_{vi}. If person v solves item i correctly, we denote that $A_{vi}=1$; if the person fails the item, we denote that $A_{vi}=0$. We also assume that each person has a certain level of ability which, for person v, we denote θ_v. It is assumed that θ_v remains stable for at least the duration of the test. We furthermore assume that each item has a certain level of easiness (i.e., the inverse of difficulty) which we refer to as ε_i.

We now can express in formal terms our assumption about how person ability and item easiness relate to the probability of answering an item correctly. We assume that:

$$(1) \quad P(A_{vi}=1) = \frac{\theta_v \cdot \varepsilon_i}{1+\theta_v \cdot \varepsilon_i}$$

Thus, we assume that the probability that person v answers item i correctly is a certain multiplicative function of the two factors ability and easiness.

The function may appear complex, but in fact it is not. What it essentially does is to express mathematically the two "commonsense" assumptions mentioned earlier, namely, that a person with a higher level of ability should have a

higher probability of answering an item correctly than a person with a lower level of ability and that the probability of answering an easier item correctly should be higher than the probability of answering a more difficult item correctly.

This can be seen if we put some numerical values into (1). For example, if we assume that the easiness parameter is 1 for a certain item, $P(A_{vi}=1) = 0.50$ if $\theta_v = 1$; if $\theta_v = 3$, then $P(A_{vi}=1) = 0.75$. If we assume that $\theta_v = 2$, we predict that for an item with $\varepsilon_i = 0.5$, $P(A_{vi}=1) = 0.50$, and for an item with $\varepsilon_i = 2$, we expect $P(A_{vi}=1) = 0.80$.

If the parameters attain values close to 0, the predicted value of $P(A_{vi}=1)$ also is close to 0. If the parameters attain high values, the expected value of $P(A_{vi}=1)$ is close to 1, as is required by the fact that a probability must lie between 0 and 1.

Thus, formula (1) is a convenient way of expressing mathematically what we have already expressed verbally. What is notable, however, is that (1) states explicitly that the probability of a correct answer shall be a function of item easiness and person ability only. Thus, no other factor may influence performance systematically, and in this sense (1) expresses a strong assumption, which may be wrong for a given set of data. Until proven wrong, however, the simplicity of (1) makes it extremely useful.

We can, of course, also easily determine the probability of an incorrect answer:

$$(2)\quad P(A_{vi}=0) = 1-P(A_{vi}=1) = 1- \frac{\theta_v \cdot \varepsilon_i}{1+\theta_v \cdot \varepsilon_i} = \frac{1}{1+\theta_v \cdot \varepsilon_i}$$

However, the expressions (1) and (2) do not solve our problem, since they presuppose that we know the item easiness and person ability values. Our task is to arrive at a method for finding out from empirical data what these values are. Of course, we can never find the true values of these parameters, since that is impossible in any statistical model, but we can try to derive estimates of the parameters which are as close as possible and which, with a sufficiently large sample of persons and items, are indistinguishable from the true parameter values.

Estimating the item parameters

There are many ways in which methods for estimating parameters in a model can be derived, but we will proceed in an intuitive, nonformal manner. Let's start with the very simple case in which one person with ability θ_v has answered three items, i,j,k, with easiness parameters ε_i, ε_j, and ε_k.

From (1) and (2) we can compute the probability of a correct or an incorrect answer to each of the items. However, to be able to determine the probabilities for patterns of responses of correct and incorrect answers to a set of items, we must make an assumption about stochastic independence. In this case, the assumption of stochastic independence means essentially that the probability of a correct answer to an item should not be influenced by whether the person passes or fails the other items in the test. Formally, this can be expressed in the following way for two items, i and j:

$$(3) \qquad P(A_{vi}=1,\ A_{vj}=1) = P(A_{vi}=1)\ \ P(A_{vj}=1)$$

That is, the probability of answering both items correctly is under the

assumption of stochastic independence given by the product of the probabilities of answering each of the items correctly.

The fact that stochastic independence is assumed does not imply that performance on the items is assumed to be uncorrelated in a group of persons: Persons with a high level of ability tend to get many items right, and persons with a low level of ability tend to get few items right, so if there are differences in levels of ability among the persons in a sample, we observe correlations between performance on the items. What the assumption of stochastic independence says is that at a given level of ability, $_v$ (i.e., when ability is "p..tialled out"), there should not be any correlations between performance on different items.

On three items the person can, of course, obtain anything between zero and three correct answers (r_v). For r_v=1 and r_v=2, several different response patterns are possible. By using (1) and (2) we can, under the assumption of stochastic independence, compute the probabilities of all these response patterns:

(4)

r_v	A_{vi}	A_{vj}	A_{vk}	P
0	0	0	0	$\frac{1}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_k}=\frac{1}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$
1	1	0	0	$\frac{\theta_v\cdot\varepsilon_i}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_k}=\frac{\theta_v\cdot\varepsilon_i}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$
1	0	1	0	$\frac{1}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{\theta_v\cdot\varepsilon_j}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_k}=\frac{\theta_v\cdot\varepsilon_j}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$
1	0	0	1	$\frac{1}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{\theta_v\cdot\varepsilon_k}{1+\theta_v\cdot\varepsilon_k}=\frac{\theta_v\cdot\varepsilon_k}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$
2	1	1	0	$\frac{\theta_v\cdot\varepsilon_i}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{\theta_v\cdot\varepsilon_j}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_k}=\frac{\theta_v^2\cdot\varepsilon_i\cdot\varepsilon_j}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$
2	1	0	1	$\frac{\theta_v\cdot\varepsilon_i}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{1}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{\theta_v\cdot\varepsilon_k}{1+\theta_v\cdot\varepsilon_k}=\frac{\theta_v^2\cdot\varepsilon_i\cdot\varepsilon_k}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$
2	0	1	1	$\frac{1}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{\theta_v\cdot\varepsilon_j}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{\theta_v\cdot\varepsilon_k}{1+\theta_v\cdot\varepsilon_k}=\frac{\theta_v^2\cdot\varepsilon_j\cdot\varepsilon_k}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$
3	1	1	1	$\frac{\theta_v\cdot\varepsilon_i}{1+\theta_v\cdot\varepsilon_i}\cdot\frac{\theta_v\cdot\varepsilon_j}{1+\theta_v\cdot\varepsilon_j}\cdot\frac{\theta_v\cdot\varepsilon_k}{1+\theta_v\cdot\varepsilon_k}=\frac{\theta_v^3\cdot\varepsilon_i\cdot\varepsilon_j\cdot\varepsilon_k}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$

Let's take a closer look at the three response patterns with $r_v=1$. If we sum the probabilities for these three patterns, we get the person's probability of obtaining a score of 1 on these items, since there is no other way in which a score of 1 can be obtained. Thus:

$$(5)\quad P(r_v=1)=\frac{\theta_v\cdot\varepsilon_i+\theta_v\cdot\varepsilon_j+\theta_v\cdot\varepsilon_k}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}=\frac{\theta_v(\varepsilon_i+\varepsilon_j+\varepsilon_k)}{(1+\theta_v\cdot\varepsilon_i)(1+\theta_v\cdot\varepsilon_j)(1+\theta_v\cdot\varepsilon_k)}$$

We next determine the conditional probability of obtaining a correct answer on item i, given that a score of 1 has been obtained (Π_{1i}). The conditional

probability of an event A, given that an event B has happened, is written $P(A|B)$ and is defined $P(A|B) = \frac{P(AB)}{P(B)}$. We thus get:

$$(6) \quad \pi_{1i} = \frac{\frac{\theta_v \cdot \varepsilon_i}{(1+\theta_v \cdot \varepsilon_i)(1+\theta_v \cdot \varepsilon_j)(1+\theta_v \cdot \varepsilon_k)}}{\frac{\theta_v(\varepsilon_i+\varepsilon_j+\varepsilon_k)}{(1+\theta_v \cdot \varepsilon_i)(1+\theta_v \cdot \varepsilon_j)(1+\theta_v \cdot \varepsilon_k)}} = \frac{\varepsilon_i}{\varepsilon_i+\varepsilon_j+\varepsilon_k}$$

In the same way we get for items j and k:

$$\pi_{1j} = \frac{\varepsilon_j}{\varepsilon_i+\varepsilon_j+\varepsilon_k} \quad ,$$

(7)

$$\pi_{1k} = \frac{\varepsilon_k}{\varepsilon_i+\varepsilon_j+\varepsilon_k} \quad .$$

Interestingly enough, we find that the ability parameter disappears from these conditional probabilities. Thus, for each and every person who obtains a raw score of 1, these conditional probabilities are the same.

The fact that the ability parameter does not appear at all in these conditional probabilities is characteristic of the Rasch model, and it has an important implication for estimation: We can estimate the item parameters without estimating the person parameters.

The three conditional probabilities above predict proportions of correct answers to each of the items among the persons who have one correct answer. Therefore, we can set up a system of equations and solve for the unknown parameters.

Let's do that for the data presented in the Introduction. Among the 4,083 persons in Sample 1 who had a raw score of 1, we observed the following frequencies and proportions of correct answers:

Item	Frequency	Proportion
1	3,567	0.874
2	456	0.112
3	60	0.015

We thus can set up the following equations:

$$(8) \quad \begin{cases} \dfrac{\varepsilon_1}{\varepsilon_1+\varepsilon_2+\varepsilon_3} = .874 \\ \dfrac{\varepsilon_2}{\varepsilon_1+\varepsilon_2+\varepsilon_3} = .112 \\ \dfrac{\varepsilon_3}{\varepsilon_1+\varepsilon_2+\varepsilon_3} = .015 \end{cases}$$

We have three unknown item parameters and we have three equations. Unfortunately, however, there is another constraint imposed on the proportions, namely, that they always sum to 1. We therefore have only two independent restrictions imposed on the item parameters and can determine only two of them. One way to solve this problem is to put the parameter value for one of the items equal to 1. In this way we do not determine the absolute values of the item parameters, but the easiness of the items relative to one of them. If we set $\varepsilon_2=1$, we get the following equations:

$$(9)\quad \begin{cases} \dfrac{\varepsilon_1}{\varepsilon_1+\varepsilon_3+1} = .874 \\ \dfrac{1}{\varepsilon_1+\varepsilon_3+1} = .112 \\ \dfrac{\varepsilon_3}{\varepsilon_1+\varepsilon_3+1} = .015 \end{cases}$$

By dividing the first equation with the second, we get $\varepsilon_1 = 7.80$, and by dividing the third equation with the second, we get $\varepsilon_3 = 0.13$. Thus, item 1 is about eight times as easy as item 2, and item 2 is about eight times as easy as item 3.

So far, we have been able to get at least some kind of estimate of the relative easiness of the three items given to Sample 1. But we have used only a part of our data, and for our procedure to be entirely satisfactory, we should use the entire set of observed persons. However, in order for us to use the other observations, they must yield essentially the same estimates. We should, therefore, convince ourselves that, for the persons who scored 2 on the test, we can get the same estimates, within statistical limits, as we obtained for those who scored 1.

The probability of obtaining a score of 2 is, of course, given by the sum of the probabilities for each of the response patterns with raw score 2 (cf. 4), i.e.:

$$(10)\quad P(r_v=2) = \frac{\theta_v^2\varepsilon_i\varepsilon_j+\theta_v^2\varepsilon_i\varepsilon_k+\theta_v^2\varepsilon_j\varepsilon_k}{(1+\theta_v\varepsilon_i)(1+\theta_v\varepsilon_j)(1+\theta_v\varepsilon_k)} = \frac{\theta_v^2(\varepsilon_i\varepsilon_j+\varepsilon_i\varepsilon_k+\varepsilon_j\varepsilon_k)}{(1+\theta_v\varepsilon_i)(1+\theta_v\varepsilon_j)(1+\theta_v\varepsilon_k)}$$

The probability of a correct answer to item i is given by the sum of the probabilities of the response patterns which include a correct response to item i, and we get the conditional probability, given $r_v=2$, as:

$$(11)\quad \pi_{2i} = \frac{\dfrac{\theta_v^2\varepsilon_i\varepsilon_j+\theta_v^2\varepsilon_i\varepsilon_k}{(1+\theta_v\varepsilon_i)(1+\theta_v\varepsilon_j)(1+\theta_v\varepsilon_k)}}{\dfrac{\theta_v^2(\varepsilon_i\varepsilon_j+\varepsilon_i\varepsilon_k+\varepsilon_j\varepsilon_k)}{(1+\theta_v\varepsilon_i)(1+\theta_v\varepsilon_j)(1+\theta_v\varepsilon_k)}} = \frac{\varepsilon_i\varepsilon_j+\varepsilon_i\varepsilon_k}{\varepsilon_i\varepsilon_j+\varepsilon_i\varepsilon_k+\varepsilon_j\varepsilon_k}$$

For items j and k we similarly get:

$$(12)\quad \pi_{2j} = \frac{\varepsilon_i\varepsilon_j+\varepsilon_j\varepsilon_k}{\varepsilon_i\varepsilon_j+\varepsilon_i\varepsilon_k+\varepsilon_j\varepsilon_k}$$

$$\pi_{2k} = \frac{\varepsilon_i\varepsilon_k+\varepsilon_j\varepsilon_k}{\varepsilon_i\varepsilon_j+\varepsilon_i\varepsilon_k+\varepsilon_j\varepsilon_k}$$

In the observed data for Sample 1, there were 4,017 persons who had a score of 2, and for these persons the following frequencies and proportions of correct answers were observed:

Item	Frequency	Proportion
1	3,955	0.985
2	3,531	0.879
3	548	0.136

Again, we can construct a system of equations by setting equal the conditional probabilities of a correct answer to each of the items and the proportions of correct answers to the items for the persons with an observed score of 2:

$$(13)\qquad \begin{cases} \dfrac{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3}{3_1{}^3{}_2+3_1{}^3{}_3+3_2{}^3{}_3} = .985 \\[2ex] \dfrac{\varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3}{\varepsilon_1\varepsilon_2\ \varepsilon_1\varepsilon_3\ \varepsilon_2\varepsilon_3} = .879 \\[2ex] \dfrac{\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3}{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3} = .136 \end{cases}$$

As before, we have three unknowns, but again there are only two independent restrictions since the proportions always sum to 2. We solve this problem in the same way as before, i.e., by setting the parameter value for item 2 equal to 1. We can then solve the equations for the two remaining item parameters. We get the results: ε_1=7.84 and ε_3=0.14. These values are very close to those obtained for score group 1, and the small variations we observe can be accounted for by stochastic factors which inevitably come into play even in a sample as large as this one.

Thus, we have shown that it is possible to get highly similar estimates of the relative easiness of items from groups of persons who have a different number of correct answers. But what about the two remaining score groups, i.e., those with a score of 0 and those with a score of 3? Unfortunately, these are quite useless for the purpose of estimating the item parameters, which is seen from the fact that in these groups the predicted proportion of correct answers is the same for all items. Thus, we have to exclude from consideration those persons who have obtained scores of 0 or 3 on the test.

Even though it has been shown that we can use either score group 1 or score group 2 and still get the same results, our estimation procedure is not yet entirely satisfactory: We would like to employ simultaneously information from both these groups of persons to get one common set of estimates.

An easy way of doing that is to put all the information into one set of equations. In doing so, however, we should not use the proportions of correct answers but the frequencies, so that the score groups influence the result in relation to their size. What we predict is then the total number of correct answers to each item, which we call s_i. If we refer to the number of persons with r correct answers as n, we can write the equation for item i:

$$(14) \quad s_i = \sum_{r=1}^{k-1} n_r \pi_{ri}$$

If we now replace π_{ri} with the corresponding expressions in terms of the item parameters (cf. 6, 7, 11, and 12) which we have previously derived, we get for the three-item test given to Sample 1 the following equations:

$$(15) \quad \begin{cases} \dfrac{4083\ \varepsilon_1}{\varepsilon_1+\varepsilon_2+\varepsilon_3} + \dfrac{4017(\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3)}{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3} = 3567+3955 \\ \dfrac{4083\ \varepsilon_2}{\varepsilon_1+\varepsilon_2+\varepsilon_3} + \dfrac{4017(\varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3)}{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3} = 456 + 3531 \\ \dfrac{4083\ \varepsilon_3}{\varepsilon_1+\varepsilon_2+\varepsilon_3} + \dfrac{4017(\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3)}{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3)} = 60 + 548 \end{cases}$$

As before, there is a constraint imposed on the system since $rn_r=\Sigma_{si}$, and we put equal to 1. After a series of algebraic manipulations, we can solve the equations and we get the result: ε_1 =7.862 and ε_3 =0.139. Again we find that these estimates are close to those we have obtained earlier, as might be expected from the quite large size of the sample.

We now have arrived at a procedure for obtaining estimates of the item parameters. This procedure yields so-called maximum likelihood estimates, and since it is based on conditional expectations, it is referred to as a conditional maximum likelihood (CML) procedure. We will not go into any depth here about different estimation procedures (see instead, Fischer 1974; Wright and Douglas 1977); suffice it to point out that the CML estimates have very good properties and that it is a general characteristic of the kinds of models with which we are dealing here that maximum likelihood estimates are obtained if observed sample characteristics are predicted from model parameters (cf. Andersen 1980).

Even in our very simple three-item formulation, the mathematical expressions involved in the estimation procedure are quite complex and, with just a few more items, they would fill several pages. Therefore, we need a simpler notation for a general formulation of the estimation equations.

The complexity stems from the sums of products of the item parameters which appear in the conditional probabilities (i.e., in the π_{ri}). One of these appears in the numerator (i.e., "above the line"), the other in the denominator (cf. 6, 7, 11, and 12).

The latter involves, for score r, the sum of all possible products of the item parameters taken r at a time. For example, for r=1 we had (cf. 6):

(16) $\varepsilon_1+\varepsilon_2+\varepsilon_3$

and for r=2 we got (cf. 11):

(17) $\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3$

These expressions are called the elementary symmetric functions of order r in the item parameters and we will denote them γ_r. γ_0 is defined to be 1. In the numerator of π_{ri} we had for item 1 and for r=1 the following (cf. 6):

(18) ε_1

For r=2 we had (cf. 11):

(19) $\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3$

These sums of products are related to the γ_r in a relatively simple way. If from γ_r we take away ε_i by crossing out all those products in which ε_i does not appear and by crossing out ε_i in those products in which it does appear, we get the elementary symmetric function of order r-1 in all parameters except ε and we denote that $\gamma^{(i)}_{r-1}$. For example, from (17) we get $\gamma^{(1)}_1$ as:

(20) $\varepsilon_2+\varepsilon_3$.

If we then multiply this expression with ε_1, we get:

$$(21) \quad \varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3$$

which is the same as (19). Since the denominator of the formula for π_{21} was γ_2, we get:

$$(22) \quad \pi_{21} = \frac{\varepsilon_1\gamma_1(1)}{\gamma_2}$$

The general form of the expression is, of course:

$$(23) \quad \pi_{ri} = \frac{\varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r}$$

and we now can write the general form of the estimation equations as (cf. 14):

$$(24) \quad s_i = \sum_{r=1}^{k-1} \frac{\varepsilon_i \gamma_{r-1}^{(i)} n_r}{\gamma_r} \qquad (i=1,...,k)$$

These equations can be solved by hand only when there are two or three items, but with larger sets of items they can be solved relatively easily with the help of a computer (e.g., Fischer 1974; Gustafsson 1979a, 1980a).

Estimating the person parameters

We now have shown how parameter values reflecting the relative easiness of a set of items can be determined. But our task is only partially completed:

We have still not dealt with the person parameters (or the ability parameters), since these disappeared from our estimation equations for the item parameters.

In principle, we could proceed in the same way with the person parameters as we did with the item parameters. That is, we could determine the conditional probabilities of obtaining different raw scores, given the observed item scores, and we would then find that the item parameters disappear, leaving us with equations in the person parameters only. But we would not be able to solve these equations, because the elementary symmetric functions in the person parameters are of immense complexity even for very small samples of persons.

We will, therefore, take another approach, using the item parameters we have already determined. It has already been concluded that if predictions involving the parameters to be estimated are equated to observed sample characteristics, we can get estimates of the unknown parameters.

The basic model (1) expresses the probability of a correct answer to an item for a certain person. The sum of these probabilities gives us the expected number of correct answers on the set of items for this person, and we can put this expression, with the already estimated item parameters inserted, equal to the observed raw score, i.e.:

$$(25) \qquad r_v = \sum_{i=1}^{k} \frac{\theta_v \varepsilon_i}{1+\theta_v \varepsilon_i}$$

This expression does not take into account which particular items the person answered correctly, so all persons with the same raw score must get the same estimated person parameter. Therefore, we only need to solve (25) for each

of the 1 to k-1 different raw scores on a set of k items (for r=0 and r=k no estimates can be obtained).

We will illustrate this for raw scores 1 and 2 on our three-item test, using the item parameters estimated from the total sample. Thus, we get the following equations:

$$1 = \frac{7.86\theta_1}{1+7.86\theta_1} + \frac{1\theta_1}{1+1\theta_1} + \frac{.14\theta_1}{1+.14\theta_1}$$

and

$$2 = \frac{7.86\theta_2}{1+7.86\theta_2} + \frac{1\theta_2}{1+1\theta_2} + \frac{.14\theta_2}{1+.14\theta_2}$$

Each of the equations resulted in a polynomial and, after a sequence of algebraic manipulations, we get the following results: $\theta_1 = .32$ and $\theta_2 = 2.97$.

Interpretations of the parameters

Having shown that it is possible to separate person ability and item easiness theoretically, and also that it is possible to compute estimates of these parameters from a set of observational data, it may be asked how these parameters can be interpreted.

Let us first summarize our results. On the basis of the total sample (except, of course, those persons with zero and three correct answers) we obtained the following results:

Item	Easiness	Raw score	Ability
1	7.86	1	0.32
2	1.00	2	2.97
3	0.14		

A higher item parameter indicates an easier item. However, we may want our scale to reflect item difficulty instead, so that a higher value indicates a more difficult item. This is easily achieved if we take the inverse (i.e., 1) of the item parameters. We then get the following item parameters:

Item	Difficulty
1	0.13
2	1.00
3	7.14

If we call these difficulty parameters , we can rewrite our basic model (1) in the following way:

$$(26)\quad P(A_{vi}=1) = \frac{\frac{\theta_v}{\delta_i}}{1 + \frac{\theta_v}{\delta_i}}$$

This formulation of the model and (1) give, of course, identical results.

On the difficulty scale, the parameters can range from 0 to positive infinity. However, on such a scale, the parameters may be difficult to interpret, and especially so since the person parameters and the item parameters combine multiplicatively. To solve this problem, we can make a new transformation in which we take the natural logarithm of both the item parameters and the

person parameters. We then get the following set of parameters (observe that only two decimal values are presented but that the calculations have been carried out with greater accuracy):

Item	Log difficulty	Raw score	Log ability
1	-2.06	1	-1.14
2	0.00	2	1.09
3	1.97		

If we set $\log \delta = \sigma$ and $\log \theta_v = \xi_v$, we can reformulate (26) into:

$$(27)\quad P(A_{vi}=1) = \frac{\exp(\xi_v - \sigma_i)}{1+\exp(\xi_v - \sigma_i)}$$

Again, this formulation of the model gives results identical to (1) and (26) (granted, of course, that in each of the formulations of the model we insert the correct set of parameters).

We will make one more transformation of our scales. It will be remembered that, when the item parameters were estimated, we chose to set ε_2 equal to 1 (and, thereby $\sigma_2=0$). The fact that we cannot estimate the absolute item difficulties but only relative item difficulties is a problem that is more often solved in another way, however; namely, by imposing the constraint that the log difficulty parameters should sum to 0 (or, equivalently, that the product of the easiness parameters should be 1). However, we can easily make this transformation now. The σ_i parameters above sum to -0.09. To change this to a sum of 0, we have to add 0.03 to each of the σ_i parameters. But if we do that, we also have to add 0.03 to each of the ξ_v parameters,

since otherwise (27) will not give the same results as before the transformation. We then get the following set of parameters:

Item	σ_i	Raw score	ξ_v
1	-2.03	1	-1.11
2	0.03	2	1.12
3	2.00		

In most applications of the Rasch model, the item parameters are expressed in this way, i.e., on a log difficulty scale on which the parameters sum to 0. However, other transformations are also common, for example, so that negative numbers are avoided (see, e.g., Wright and Stone 1979).

Let us now take a look at some statements that can be made on the basis of the estimated parameters.

The odds of success to an item are defined as the ratio of the probability of success to the probability of failure, i.e.:

$$(28)\quad \lambda_{vi} = \frac{P_{vi}}{1-P_{vi}} = \frac{\frac{\exp(\xi_v-\sigma_i)}{1+\exp(\xi_v-\sigma_i)}}{\frac{1}{1+\exp(\xi_v-\sigma_i)}} = \exp(\xi_v-\sigma_i)$$

For example, if the probability of success is 0.75, the odds are 0.75/0.25=3.

The ratio of odds of success to two items, i and j, is:

$$(29)\quad \frac{\lambda_{vi}}{\lambda_{vj}} = \frac{\exp(\xi_v-\sigma_i)}{\exp(\xi_v-\sigma_j)} = \exp(\sigma_j-\sigma_i)$$

We find that this ratio is governed exclusively by the item parameters; again the person parameter disappears. This, of course, indicates that the ratio is constant over the range of ability and that we can compare items independently of persons.

The same types of statements about relative odds can also be made about persons. If we relate the odds of success for person v to the odds of success for person u on item i, we find that:

$$(30) \qquad \frac{\lambda_{vi}}{\lambda_{ui}} = \frac{\exp(\xi_v - \sigma_i)}{\exp(\xi_u - \sigma_i)} = \exp(\xi_v - \xi_u)$$

Here the item parameter disappears and we can conclude that the relative odds are invariant over items; i.e., we can compare persons independently of items. Quite interesting practical consequences follow from this: If the persons have been given different items, we still can compare their abilities, granted, of course, that we have previously determined the relative difficulties of the items.

The possibility of making comparison of persons independently of items, and of items independently of persons, forms the core of Rasch's "theory of specific objectivity" (Rasch 1960, 1977). Expressed in simple terms, this model requires that in comparing different "objects" (here persons), different "agents" (here items) should give the same rank order among the objects. That is, if we rank order the persons in a sample on the basis of their performance on one set of items, we would expect this rank ordering to be the same for another set of items which purports to measure the same dimension. If this is not the case, we would regard the two sets of items

as measuring different dimensions, and we should hesitate before using them interchangeably. The theory of specific objectivity is, therefore, basically a theory of unidimensionality; i.e., it states that only one dimension should be measured at the same time and that all the items which are used should be homogeneous in the sense that they all measure the same dimension.

Testing the Assumptions of the Rasch Model

We have already indicated some of the possibilities for practical applications of the Rasch model. However, before we proceed to treat some of these applications in greater detail, it is necessary to discuss the limits of the applicability of the Rasch model.

Even though it can be shown in theory that it is possible to separate person ability and item difficulty from the answers to a set of items, this does not always happen in practice. This is because the model is based on a series of assumptions which may not be fulfilled, and if these assumptions are violated, we cannot estimate item parameters independently of persons, and we cannot estimate ability independently of items.

Expressed briefly and in simple terms, the following assumptions are made:

1. We assume that the probability of a correct answer to an item is a simple function of a person parameter and an item parameter, as is expressed in (1).

2. The items are assumed to be homogeneous; i.e., it is assumed that all items measure the same ability.

3. Local stochastic independence is assumed; i.e., it is assumed that the response to one item does not affect the response to another item. It is also assumed that the responses made by one person do not affect the responses of another person.

It can be shown that assumption 3 formally reduces to assumption 2 (cf. Gustafsson 1980b), so we only have to deal with the first two assumptions.

The basic question to decide is whether one or both of these assumptions are not fulfilled. If that is the case, the Rasch model will not fit the data, since the model and the data are based upon incompatible assumptions.

One way to do this is, of course, to look at the test items to see whether there is anything in them which may cause a violation of the assumptions. For example, if the items can be divided into groups which clearly measure different abilities, we can be sure that the assumption of unidimensionality is being violated. If, to take another example, there is ample opportunity of guessing the correct answer, as is often the case in multiple-choice items, we also may expect that the data will not fit the model since (1) does not allow for guessing.

But however useful such procedures may be, it would, of course, be desirable if the model itself could tell, as it were, if a set of data does or does not fit.

It will be remembered that we have shown for our three-item test for Sample 1 that the relative difficulties of the items remained roughly the same whether we used persons with one correct answer or persons with two correct answers to estimate the parameters. Below are shown frequencies of correct

answers to three items in another test for persons with one and two correct answers:

	Frequency		
Item	r=1	r=2	
1	988	4,110	
2	224	3,169	n_1=1,368
3	156	1,527	n_2 =4,403

Applying the same procedures as before, we get the following estimates of the item parameters for the two groups of persons:

Item	r=1	r=2
1	-1.11	-1.24
2	0.37	0.20
3	0.74	1.04

In this case, it is obviously not true that the relative difficulties of the items remain invariant over the score groups. The reason for this is that the data analyzed here have resulted from a true-false test, in which there is ample opportunity to guess the correct answer.

Thus, we find that when the data do not fit the model, it is no longer true that the item parameters are invariant over groups of persons. To investigate this property of the model, we can thus estimate the item parameters within groups of persons with a different number of correct answers, to see whether we get the same results. When the sample of persons is small, there will, of course, be quite a variation among the item parameters due to chance factors, and it may be difficult to judge whether there are any true differences among the item parameters. Andersen (1973; cf. Gustafsson 1980b)

has, however, suggested a procedure that gives a summary statistical test of the equality of the item parameters.

Applying this test to Sample 1, we find the test statistic to be χ^2 =0.19 with 2 degrees of freedom. Obviously, we cannot reject the null hypothesis that the item parameters estimated within the two score groups are equal. For the data presented in this section, the test statistic is χ^2 =23.59. With 2 degrees of freedom this is a very highly significant value, and we must conclude that there are significant differences among the item parameters for the two score groups.

This test investigates whether the item parameters are invariant over groups of persons, and it can be applied with the sample of persons divided in any possible way. But the test does not investigate whether a set of items is unidimensional or not. To do that, we should study whether the person parameters are invariant over groups of items or not.

To study whether the person parameters are invariant over groups of items, we could, of course, divide the items into subsets suspected to measure different abilities, and then for each person study whether the person parameters estimated from the different groups of items are the same. But such a test would involve a very large number of comparisons, and it would be a very unstable test since each person parameter would, in most cases, be estimated from few items. Martin-Lof (1973; cf. Gustafsson 1980b) has, however, suggested a test of the invariance of each person parameter over groups of items in which the actual estimation of each person parameter is avoided. This test requires, however, at least two

items in each group, so we cannot illustrate its use on our data. The application of the test is quite straightforward, however.

If we have investigated the invariance of item parameters over groups of persons and the invariance of person parameters over groups of items, and if we have been able to conclude that the data fit the model, we can proceed with any application of the model we would like. However, in many cases, it is possible to carry out sensible applications of the model even though it has been necessary to conclude that one or more of the assumptions may have been violated. This is because the presence of small violations of the assumptions need not matter at all for the intended application. But if the model is to be used in the presence of such violations, it must be made likely that the violations of the assumptions do not carry any negative implications for the validity of the conclusions drawn.

APPLICATIONS OF THE RASCH MODEL

We can now return to our question in the Introduction: How do we compare scores obtained on different tests by different groups of persons?

We have already estimated the item parameters for Sample 1, and doing that for Sample 2 as well, we get the following results:

Item	Sample 1	Sample 2
1	-2.03	-1.51
2	0.03	0.00
3	2.00	1.51

It will be remembered that item 3 was common for the two tests, whereas the other items were all different. Since we have concluded that item parameters are (or can be) invariant over groups of persons, we might perhaps expect to find the same estimate of σ_3 in the two samples. But this is obviously not the case, there being a substantial difference in the two estimates of the difficulty of item 3.

The reason for this is that the difficulties of the items can only be determined relative to each other. Thus, if item 3 is administered along with two easy items, it appears to have a high difficulty, and if it is administered along with two difficult items, it appears to have a low difficulty. But we know that the true difficulty of an item is the same independently of which persons have answered it and which other items happened to be administered along with it. Therefore, we can relate the difficulties of noncommon items to each other by using common items as points of reference.

The difficulty of item 3 is 0.49 units higher in the Sample 1 test than in the Sample 2 test. This must mean that items 1 and 2 in the Sample 1 test test are easier than the corresponding items in the Sample 2 test. To translate the item difficulties of the two sets of items so that the difficulties of all items are directly comparable, we must make the transformation in such a way that the difficulty of item 3 is the same in the two tests. We must, therefore, subtract 0.49 from item 3 in the Sample 1 test, and, of course, also from the other two items in this test (we could, of course, also have added 0.49 to the items in the Sample 2 test). After this transformation we get the following item difficulties:

Item	Sample 1	Sample 2
1	-2.52	-1.51
2	-0.46	0.00
3	1.51	1.51

The item difficulties are now expressed on the same scale, and we can estimate the person parameters corresponding to different raw scores on the two tests:

Raw score	Sample 1 test	Sample 2 test
1	-1.60	-0.94
2	0.63	0.94

If we compare these person parameters corresponding to different raw scores on the Sample 1 test with those we have previously estimated, we find that there is a constant difference of 0.49.

We now could proceed to estimate a person parameter for each of the persons in the two samples. We also could compute the means and the standard

deviations of the person parameters and compare the samples with respect to these statistics. The frequency distributions of raw scores are presented below:

	Raw score			
	0	1	2	3
Sample 1:	906	4,083	4,017	994
Sample 2:	926	3,316	4,088	1,670

It is quite obvious, however, that this procedure would not give us a true picture of the difference in level of performance of the two groups of persons. The problem is that we cannot estimate the person parameters for persons with a raw score of 0 and 3, and since there are more persons with a score of 3 in Sample 2 than in Sample 1, we would underestimate the level of performance of Sample 2 if we excluded these persons.

There is a solution to this problem, however. Andersen and Madsen (1977) have shown that if the distribution of person parameters is assumed to be of a certain kind, it is not necessary to estimate each of the person parameters; instead, parameters describing the distribution can be estimated. In this procedure it is not necessary to exclude persons who have no correct answer or only correct answers, and it can be applied whenever a reasonable assumption can be made about the distribution of person parameters. Often, a normal distribution can be assumed. It is, unfortunately, impossible to treat here the details of this quite complex procedure, but with the help of a computer, it is a simple matter to determine the parameters of the assumed distribution and also to perform a statistical test of whether the assumption made is reasonable or not.

Assuming that in each of the samples the person parameters are normally distributed, the two parameters (mean and stand deviation) which describe the normal distribution have been estimated, using the difficulty parameters on the common scale:

	Sample 1	Sample 2
Mean	-0.46	0.30
SD	1.01	1.00

We find that the SD's are the same, but that the mean of Sample 2 is 0.76 units (or about 75 percent of an SD) higher than the mean of Sample 1.

The validity of this conclusion to large extent hinges on the correctness of the assumption that the person parameters are normally distributed. The statistical test of the assumption of normality gives for Sample 1 χ^2 =1.04 and for Sample 2 χ^2 =0.22. With 1 degree of freedom none of these values is significant, so we don't find any reason to doubt the correctness of the assumption of normality.

It can be noted that we are able to draw the conclusion that the distribution of person parameters is normal in spite of the fact that the distribution of raw scores, especially for Sample 2, is quite far from normal. This is, of course, because the test given to Sample 2 was somewhat too easy for this group of persons, resulting in a negatively skewed distribution of raw scores.

We now have solved the problem formulated in the Introduction to this paper, and we have also indicated the basic properties of the Rasch model. It is obvious, however, that this model lends itself to the solution of measurement problems other than this particular one, and we will now briefly indicate some of these possible areas of application.

Test linking and equating; item banks

The problem that we have dealt with is often referred to as a problem of test linking. In practical applications, this problem of estimating item parameters, and, thereby, also person parameters, on a common scale is approached in the same way as we have done it, except that more than one item is usually common between the tests. Often ten or more such items are used. When more than one common item is used, the means of the difficulties of these items are compared across the tests to determine the constant to be added or subtracted from one or more of the sets of estimated item parameters.

Sometimes one group of persons has been given two or more tests measuring the same dimension. If we then want to express raw scores obtained on different tests on the same scale, this problem of test equating is easily solved with the Rasch model: All that needs to be done is to estimate the item parameters with the items in all tests pooled and then to use these item parameters to compute the person parameters corresponding to each raw score on the different tests.

By applying repeatedly linking and equating procedures, it may be possible to determine on a common scale the item parameters for a large set of items. Such item banks may, of course, be extremely useful since it is possible to construct a virtually infinite number of test forms, which all give person parameters on the same scale, by selecting items from the bank. Special designs to optimize the creation of item banks have been developed (see e.g., Wright 1977; Wright and Stone 1979).

Test optimization; level testing and tailored testing

When a parameter in a model is estimated from data, the estimate can have varying degrees of precision. For example, if a small sample of persons is used, the estimates of the item parameters will tend to be unstable; if we estimate the item parameters anew from an equally small sample, we may observe quite large differences between the sets of estimates. Not only the number of persons and items in a sample determine the precision with which the item and the person parameters can be estimated, but also the relative size of the parameters themselves. The statistical information (I) about the parameters contributed by a response is given by the expression:

$$(31) \quad I_{iv} = P(A_{vi}=1)(1-P(A_{vi}=1)) = \frac{\exp(\xi_v-\sigma_i)}{(1+\exp(\xi_v-\sigma_i))^2}$$

I_{iv} is highest when the probability of a correct answer is 0.5. The information about the person parameter contained in the responses to k items is the sum of the information contributed by each of the responses, i.e.:

$$(32) \quad I(\xi_v) = \sum_{i=1}^{k} I_{iv}$$

The standard error of measurement of the person parameter ($SEM[\xi_v]$) can be obtained from (32). It is:

$$(33) \quad SEM(\xi_v) = \frac{1}{\sqrt{I(\xi_v)}}$$

When the number of items is large (larger than 20-30), the SEM (ξ_v) is normally distributed, which makes it possible to construct confidence intervals around the estimated person parameter.

It is quite obvious that the difficulties of the items in a test affect the SEM (ξ_v) and also that the precision with which we can estimate any person parameter is a function of the person parameter itself.

For example, if we have a test of twenty-five items all with difficulty 0, the SEM (ξ_v) for persons with an ability of 0 is 0.4. But for persons with ability -2.0 (and 2.0), the SEM (ξ_v) is 0.62. To achieve an equally low SEM (ξ_v) for these levels of ability, we would need to administer as many as sixty items.

But, of course, another possibility would be to administer twenty-five items of difficulty -2.0 to the persons who have ability around -2.0. Doing this would gain not only shorter testing time but perhaps also a greater motivation in taking the test, since it is likely to be quite frustrating to try to answer a large number of items when the probability of a correct answer is low, and quite boring to do so when the probability of a correct answer is high.

Such a procedure would illustrate the use of <u>level tests</u>, with different levels of difficulty for different groups of persons. It is necessary, of course, to have some preliminary estimate of the level of ability of the persons in order to be able to assign the different forms. Such a preliminary estimate may be obtained either from a short pretest or from results achieved on previously administered tests.

But the procedure can, of course, also be developed into a tailored testing procedure, in which, for each person, the response to one item determines which item will be given next. It is not difficult to devise such a system on the basis of the Rasch model, granted that the test items can be administered by a computer.

THE RASCH MODEL AND OTHER TEST-THEORETIC MODELS

We will end this introduction to the Rasch model by briefly indicating differences and similarities between this model and some other test-theoretic models.

Classical Test-Theory

In classical test-theory (e.g., Gulliksen 1950; Lord and Novick 1968), the concept of item difficulty has a prominent place too. The statistic most commonly used to index item difficulty is the proportion of correct answers to the item, or the p-value. But as we saw in our empirical example, the p-value varies with level of ability of the sample. This statistic, therefore, cannot be generalized from one sample of persons to another, unless they are random samples from the same population. In the Rasch model, in contrast, the estimates of item difficulty remain invariant from one sample of persons to another, at least as long as the data fit the model.

In classical test-theory, the observed number of correct answers is most commonly taken to represent person ability. Scores obtained on different sets of items are not comparable, however, unless the test forms have been carefully equalized. In the Rasch model, too, ability is estimated from the number of correct answers, but since item difficulty is taken into account, results obtained on different sets of items may be compared.

For purposes of item screening, statistics which reflect the relation between item performance and test performance are relied upon within classical test-theory. Among such measures of item discrimination, the biserial and point-biserial correlation coefficients are the ones most

frequently used. When the purpose of the test is to reflect individual differences in performance, items with a high relation to overall test performance are sought. In the Rasch model, there is no concept which corresponds to the indices of item discrimination, since it is assumed that all items have the same degree of relationship with ability. Therefore, when items are selected for inclusion in a Rasch scale, the criterion is evenness of discrimination, rather than highness of discrimination. However, nothing prevents Rasch scales to be constructed from items with high and even discrimination (cf. Gustafsson 1980b).

Reliability is the measurement concept which plays the most important part in classical theory. This concept reflects the accuracy with which a group of individuals can be rank ordered on the basis of test performance. The observed test variance is assumed to consist of two independent parts: true variance, reflecting the true individual differences in performance; and error variance, reflecting random variation. A simple definition of the coefficient of reliability is the ratio of true variance to observed variance. Since the sample variance of ability enters into the estimate of the reliability coefficient, this coefficient expresses properties of both the test and the sample of persons.

In the Rasch model the concept of reliability plays a subordinate part, because this measurement model is oriented toward estimation of individual ability, rather than toward comparison of individuals. The SEM (ξ_v) is, therefore, the most important index of the accuracy of measurement in the Rasch model.

In conclusion, we have seen a recurring difference between classical test-theory and the Rasch model; in the former there is a dependence on specific samples of persons and items, whereas in the latter this is not the case. It is obvious, therefore, that when an application requires measurement of the same dimension with different sets of items, the Rasch model offers great advantages. In other, less demanding measurement applications, however, the greater complexity and stricter assumptions of the Rasch model may make classical test-theory a good alternative.

Other item-response models

The Rasch model may be viewed as one member of a larger family of models, all members of which have the characteristic that an explicit model is specified for the relation between observable item performance and an unobservable trait assumed to underlie performance. These item-response, or latent-trait, models all allow estimation of invariant person and item parameters.

The Rasch model is the simplest of these models in the sense that only one parameter, the difficulty, is used to specify item characteristics. In other models, more parameters are used to characterize the items. Thus, in the so called three-parameter model (Lord and Novick 1968), each item is described by three parameters: difficulty, discrimination, and guessing. The difficulty parameter has an interpretation similiar to that in the Rasch model. The discrimination parameter represents the degree of relationship between item performance and ability, and the guessing parameter represents the expected level of performance on the item for persons with very low ability.

This model thus makes less strict assumptions of the nature of the observations than does the Rasch model, and it can be expected to fit a wider range of empirical data than the Rasch model. There is a price to be paid for this generality, however: the three-parameter lacks the conceptual simplicity and elegance of the Rasch model; estimation of parameters in the model presents great technical difficulties and requires large samples of persons and items; and the model is considerably more difficult to apply than is the Rasch model.

There has been a rather heated debate on the relative virtues of different item-response models, some arguing that the Rasch model is always the proper choice, and others arguing that the Rasch model may never be expected to fit data. Such extreme positions seem unnecessary, however, since it appears that each of the models has both strengths and weaknesses.

The major weakness of the Rasch model is that it involves such strong assumptions. In particular, it cannot be expected to fit multiple-choice items. For some applications, such as equating tests of widely differing difficulty, violation of the assumption of no guessing may have very serious implications (Gustafsson 1979b), so great care must be exercised if the Rasch model is to be applied to items of the multiple-choice type.

The major weakness of the three-parameter model is that results obtained with this model are not dependable unless the samples of items and persons are large. This model is, therefore, best suited for large-scale applications of multiple-choice tests, whereas the Rasch model seems best suited for small- and large-scale applications in which guessing is not a major factor in test performance.

REFERENCES

Andersen, E. B. (1973) A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Andersen, E. B. (1980) Discrete statistical models with social and science applications. Amsterdam: North-Holland Publishing Company.

Andersen, E., and Madsen, M. (1977) Estimating the parameters of the latent population distribution. Psychometrika, 42, 357-374.

Fischer, G. H. (1974) Einfuhrung in die Theorie psychologischer Tests. Grundlagen and Anwendungen. Bern: Huber.

Gulliksen, H. D. (1950) Theory of mental tests. New York: Wiley.

Gustafsson, J. E. (1979a) PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items. Reports from the Institute of Education, University of Goteborg, no. 85.

Gustafsson, J. E. (1979b) The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, 16, 153-158.

Gustafsson, J. E. (1980a) A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. Educational and Psychological Measurement, 40, 377-385.

Gustafsson, J. E. (1980b) Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205-233.

Lord, F. M., and Novick, M. R. (1968) Statistical theories of mental test scores. Reading: Addison-Wesley.

Martin-Lof, P. (1973) Statistiska modeller. Anteckningar fran seminarier lasaret 1969-70 utarbetade av Rolf Sundberg. 2:a uppl. (Statistical models. Notes from seminars 1969-70 by Rolf Sundberg. 2nd ed.). Institutet for forsakringsmatematik och matematik vid Stockholms universitet.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute for Educational Research.

Rasch, G. (1977) On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements. In Danish Yearbook of Philosophy, Vol 14. Copenhagen: Munksgaard, pp. 58-94.

Wright, B. D. (1977) Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D., and Douglas, G. A. (1977) Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement, 37, 573-586.

Wright, B. D., and Stone, M. H. (1979) Best test design. Chicago: Mesa Press.